KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 3 урок > Uplift сложных метрик. Expected value framework

> Оглавление

> Методология Uplift-моделирования

Эта лекция посвящена построению uplift-моделей и expected value framework (оцениванию математического ожидания целевой величины для uplift.

Очень часто не совсем понятно, как поставленные бизнес-задачи связаны с методами и метриками машинного обучения. Как правило, задача сформулирована в неудобном для применения data science виде, а также нет прямого DS инструмента, решающего поставленную задачу. Зачастую это связано

с тем, что цели и бизнес-метрики для оптимизации сложно устроены и у бизнес-процесса есть ряд ограничений, которые нужно соблюдать.

В этой лекции мы как раз и поговорим о том, как с помощью выписывания математического ожидания бизнес-метрики придумать схему решения поставленной бизнес-задачи. Мы разберём этот подход на примере двух кейсов:

- Рассылка писем для сбора пожертвований в благотворительный фонд;
- Кампания с бонусными экспресс-баллами в Пятёрочке.

> Кейс 1: Сбор пожертвований

Задача состоит в том, чтобы правильно выбрать людей, которым мы будем отправлять письма или SMS-сообщения с предложением внести пожертвование в благотворительный фонд. Разумеется, мы хотим собрать как можно больше, потратив на это минимум средств.

Задачу можно решать несколькими способами, однако далеко не все из них оптимальны. Вот некоторые из "наивных" подходов:

- Максимизировать количество откликнувшихся. Проблема: Все люди могут пожертвовать очень маленькую сумму: например, по 1$. А ведь могут быть жертвователи, которые готовы заплатить 1000$, что не просто намного больше, но и гораздо выгоднее: чтобы получить ту же сумму, мы тратим деньги не на 1000, а только на 1 SMS-сообщение.
- Максимизировать сумму всех пожертвований. Проблема: Если сумма пожертвований каждого из откликнувшихся людей будет близка к затратам на информирование (SMS-сообщение), то выгода от кампании будет маленькой.

Разумным предложением в таком случае является оптимизация прибыли от кампании. Давайте сформируем табличку с условием задачи, чтобы проанализировать, как мы могли бы её решать:

Целевая бизнес-метрика	Прибыль = Сумма пожертвований − затраты на информирование
Ограничения задачи	Не более K человек в рассылке
На что мы влияем	Каждому человеку мы либо отправляем письмо, либо НЕ отправляем

Контекст в момент принятия решения	Описание человека (пол, возраст, история пожертвований и т.д.)
Дополнительные данные	Данные об истории прошлых рассылок

Введём следующие обозначения:

Y_{total} — прибыль от рассылки (кампании).

S_{total} — сумма собранных пожертвований.

c — стоимость одного письма/SMS-сообщения.

M — количество людей, которым отправили сообщение с предложением $(M \leq K)$.

Тогда прибыль $Y_{total} = S_{total} - cM$, а наша цель — максимизировать $Y_{total} \to \max$ с помощью правильного выбора аудитории для рассылки. Тут необходимо понимать, что Y_{total} — случайная величина (как и обычно все целевые переменные в uplift-моделировании), поэтому мы будем оптимизировать её математическое ожидание: $\mathbb{E}[Y_{total}|audience] \to \max$

Вспомним уже привычные нам обозначения из предыдущих лекций и перепишем с их помощью математическую постановку задачи:

X — признаки человека.

$T \in 0, 1$ — флаг рассылки (отправки письма/SMS-сообщения).

S — сумма пожертвования от человека (равна 0, если человек не откликнулся на призыв, либо вовсе не попал в рассылку).

$R = S > 0$ — событие "Человек что-то пожертвовал".

$Y \overset{def}{=} (S - c)T$ — "Прибыль" от человека.

Тогда суммарная прибыль будет выражаться как $Y_{total} = \sum_{i \in persons} Y_i$, а значит

$$\mathbb{E}[Y_{total}|audience] = \sum_{i \in persons} \mathbb{E}[Y|X = x_i, T = t_i]$$

Если внимательно посмотреть на эту формулу и проанализировать её, можно заметить, что $Y > 0$ только если у нас $T = 1$, а значит мы фактически получим прибыль Y, только если отправим письмо. Поэтому логичным решением будет прогнозировать факт того, что $\mathbb{E}[Y|X = x_i, T = 1] > 0$

Однако на самом деле всё ещё сложнее: в текущей реализации мы можем наткнуться на ситуацию с жертвователями в 1$. Поэтому было бы здорово сравнивать ещё и матожидания $\mathbb{E}[Y|X = x_i, T = 1]$ и $\mathbb{E}[Y|X = x_j, T = 1]$, т.е. фактически выбирать того, кто больше жертвует. Поэтому давайте будем предсказывать просто значение $\mathbb{E}[Y|X = x, T = 1]$.

Для того чтобы далее разобраться с задачей, нужно повнимательнее посмотреть на нашу целевую переменную. Справедливо следующее разложение:

$$\mathbb{E}[Y = S - c\,|X = x_i, T = 1] = \mathbb{E}[S|X = x_i, T = 1] - c = \mathbb{P}(R|X = x_i, T = 1)\,\mathbb{E}[S|R, X = x_i, T = 1] + \mathbb{P}(\bar{R}|X = x_i, T = 1) * 0 - c = \mathbb{P}(R|X = x_i, T = 1)\,\mathbb{E}[S|R, X = x_i, T = 1] - c$$

Обратите внимание, что формула даёт разложение на слагаемые/множители, которые представляют собой вероятность или матожидание какой-то "простой" величины, которую гораздо проще моделировать. Сделав такое разложение, мы понимаем:

- Какие показатели необходимо прогнозировать;
- Какие данные нужны для построения соответствующих моделей;
- Как "собрать" итоговый прогноз и какое решение принимать на его основе.

Исходя из разложенной формулы, получаем следующую схему решения:

1. Прогнозируем $\mathbb{P}(R|X = x_i, T = 1)$ и $\mathbb{E}[S|R, X = x_i, T = 1]$ с помощью машинного обучения.
2. Собираем итоговый прогноз по разложенной формуле.
3. Сортируем респондентов по прогнозу в порядке убывания.
4. Берём топ-K среди респондентов с положительным прогнозом для отправки рассылки.

> Кейс 2: Экспресс-баллы в Пятёрочке

Мы проводим промоакцию в магазине, стимулирующую покупателей Пятёрочки прийти в магазин и сделать покупку. Для этого мы рассылаем письмо/SMS-сообщение держателям бонусных карт с информацией, что им начислено N экспресс-баллов, которые можно потратить до определённой даты (обычно

неделя). После этой даты баллы сгорят. Поэтому мы хотели бы определить тех клиентов сети, разослав письма которым мы получим максимальную выручку по результатам промокампании.

Как и в прошлом кейсе, составим таблицу с информацией о задаче:

Детали промокампании	Предложение клиенту: "Вам начислено N баллов. Успейте списать их в течение следующей недели!"
Целевая бизнес-метрика	Прибыль = Сумма покупок − Себестоимость товаров − Затраты на информирование − Затраты на экспресс-баллы
Ограничения задачи	Не более K клиентов в рассылке
На что мы влияем	Решаем, делать ли клиенту акционное предложение в виде экспресс-баллов
Контекст в момент принятия решения	Описание клиента (пол, возраст, город, история покупок и т.д.)
Дополнительные данные	История прошлых промокампаний

Как и в прошлый раз, давайте перейдём к математической постановке задачи с учётом привычных обозначений (напомним их снова ниже):

X — признаки клиента.

c — стоимость SMS.

$T \in 0, 1$ — флаг отправки SMS.

$R \in 0, 1$ — флаг того, что клиент воспользовался баллами.

b — размер вознаграждения в рублях.

Z — маржа (выручка - себестоимость) покупок клиента в период кампании.

Q — число покупок клиента в период кампании.

S — средняя маржа чека в период кампании (т.е. $S = Z/Q$).

С учётом обозначений прибыль от клиента равна $Y = Z - b\,R - c\,T$

Теперь необходимо понять, что же нам нужно прогнозировать. Поскольку для каждого клиента нам нужно выбрать, делать ли ему предложение, то в случае предложения мы заработаем $\mathbb{E}[Y|X, T = 1]$, а в случае его отсутствия — $\mathbb{E}[Y|X, T = 0]$. В таком случае хорошей идей является прогнозирование того,

будет ли выгодно предложить клиенту экспресс-баллы, т.е. является ли $\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0] > 0$

Для удобства введём обозначение $Up[Y|X] \overset{def}{=} \mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0]$

Внимательный студент тут же воскликнет: "Погодите! Но ведь в предыдущем кейсе мы были в такой же ситуации, где могли ещё сравнить выгоду от предложения одного клиенту с выгодой от предложения другому клиенту. Там мы отказались от прогнозирования факта выгодности и прогнозировали саму выгоду!" И на самом деле студент будет прав! Это замечание справедливо и в этой задаче, поэтому мы будем прогнозировать сам $Up[Y|X]$.

Но прежде сделаем аналогичное разложение, как и в предыдущем кейсе:

$$Up[Y = Z - b\,R - c\,T|X] = Up[Z|X] - b\,\mathbb{P}(R = 1|X, T = 1) - c$$

Из формулы видно, что моделировать можно отдельно $Up[Z|X]$ и $\mathbb{P}(R = 1|X, T = 1)$

На самом деле $Up[Z|X]$ можно раскладывать ещё дальше на более малые составляющие. Такой анализ часто даёт возможность интерпретировать, из чего конкретно (какого явления) складывается Uplift: чаще ли люди стали ходить в магазин, выросла ли средняя сумма чека и т.д. Достаточно подробно это разложение было рассмотрено в практической части лекции (см. видео). Здесь мы его не приводим, потому что разложение сильно опирается на данные, с которыми работает лектор, а значит, эту информацию сложно генерализировать.

Что же касается прогнозирования $\mathbb{P}(R = 1|X, T = 1)$, то разумным, хотя и не очень честным, является следующее разложение:

$$\mathbb{P}(R = 1|X, T = 1) \approx \mathbb{P}(R = 1|T = 1, Z > 0)\ \mathbb{P}(Z > 0|X, T = 1)$$

Как видим, первое слагаемое вообще не учитывает данные о клиенте, а значит, его можно взять просто как среднюю вероятность списать полученные баллы при условии, что клиент совершал покупки. Для оценивания второго слагаемого можно построить модель или вообще использовать статистические эвристики наподобие доли окон в 7 дней, в течение которых клиент совершал покупки (разумеется, это будет являться очень грубой оценкой).

Иногда грубое оценивание тех или иных величин оправданно, но у такого подхода есть недостаток — он плохо адаптируется к изменениям в среде (данных): если со

временем бóльшую роль начнут играть другие слагаемые, то такой подход не подстроится к этому изменению.

> Резюме

Мы разобрали два кейса о построении Uplift-моделей для сложных метрик. Мы также познакомились с подходом оценивания математического ожидания целевой величины и разложения этого матожидания на более простые составляющие, которые гораздо легче и надежнее (робастнее) прогнозировать с помощью методов машинного обучения.

Благодаря разбору этих задач мы поняли:

1. На что следует обратить внимание при исследовании данных;
2. Какие показатели нужно прогнозировать;
3. Какие данные нужны для построения соответствующих моделей (а следовательно, какие данные могут понадобиться в будущем, т.е. это позволило нам строить планы для их генерации или сбора);
4. Как собрать итоговый прогноз и какое принять решение.

Итоговый процесс (pipeline) обычно выглядит так: